Chapter 24

CoS Overview

For interfaces that carry IPv4, IPv6, or MPLS traffic, you can configure JUNOS class-of-service (CoS) features to provide multiple classes of service for different applications. On the router, you can configure multiple forwarding classes for transmitting packets, define which packets are placed into each output queue, schedule the transmission service level for each queue, and manage congestion using a Random Early Detection (RED) algorithm.

> **Note**
>
> JUNOS CoS features are not supported on ATM interfaces. ATM has traffic-shaping capabilities that would override CoS, because ATM traffic shaping is performed at the ATM layer and CoS is performed at the IP layer. For more information about ATM traffic shaping, see "Define the ATM Traffic-Shaping Profile" on page 179.

The JUNOS CoS features provide a set of mechanisms that you can use to provide differentiated services when best-effort traffic delivery is insufficient. In designing CoS applications, you must give careful consideration to your service needs, and you must thoroughly plan and design your CoS configuration to ensure consistency across all routers in a CoS domain. You must also consider all the routers and other networking equipment in the CoS domain to ensure interoperability among all equipment.

The Internet community has little experience with CoS and quality of service (QoS). However, because Juniper Networks routers implement CoS in hardware rather than in software, you can experiment with and deploy CoS features without adversely affecting packet forwarding and routing performance.

The standards are defined in the following RFCs:

- RFC 2474, *Definition of the Differentiated Services Field in the IPv4 and IPv6 Headers*
- RFC 2598, *An Expedited Forwarding PHB*
- RFC 2597, *Assured Forwarding PHB Group*

This chapter discusses the following topics:

CoS Applications

CoS mechanisms are useful for two broad classes of applications. These applications can be referred to as *in the bo x* and *across the netw ork*.

In-the-box applications use CoS mechanisms to provide special treatment for packets passing through a single node on the network. You can monitor the incoming traffic on each interface, using CoS to provide preferred service to some interfaces (that is, to some customers) while limiting the service provided to other interfaces. You can also filter outgoing traffic by the packet's destination, thus providing preferred service to some destinations.

Across-the-network applications use CoS mechanisms to provide differentiated treatment to different classes of packets across a set of nodes in a network. In these types of applications, you typically control the ingress and egress routers to a routing domain and all the routers within the domain. You can use JUNOS CoS features to modify packets traveling through the domain to indicate the packet's priority across the domain. Specifically, you modify the precedence bits in the IPv4 type-of-service (ToS) field, remapping these bits to values that correspond to levels of service. When all routers in the domain are configured to associate the precedence bits with specific service levels, packets traveling across the domain receive the same level of service from the ingress point to the egress point. For CoS to work in this case, the mapping between the precedence bits and service levels must be identical across all routers in the domain.

JUNOS CoS applications support the following range of mechanisms:

Differentiated Services—The CoS application supports DiffServ as well as six-bit IP header ToS byte settings. The configuration uses DiffServ Code Points (DSCPs) in the IP ToS field to determine the forwarding class associated with each packet.

Layer 2 to Layer 3 CoS Mapping—The CoS application supports mapping of Layer 2 (IEEE 802.1p) packet headers to router forwarding class and loss-priority values.

Layer 2 to Layer 3 CoS mapping involves setting the forwarding class and loss priority based on information in the Layer 2 header. Output involves mapping the forwarding class and loss priority to a Layer 2-specific marking. You can mark the Layer 2 and Layer 3 headers simultaneously.

MPLS EXP—Supports configuration of mapping of MPLS experimental (EXP) bit settings to router forwarding classes and vice versa.

VPN Outer Label Marking—Supports setting of outer label EXP bits based on MPLS EXP mapping.

JUNOS CoS Components

You can configure CoS features to meet your application needs. Because the components are generic, you can use a single CoS configuration syntax across multiple platforms. The JUNOS CoS features include:

Classifiers—Allow you to associate incoming packets with a forwarding class and loss priority and, based on the associated forwarding class, assign packets to output queues. Two general types of classifiers are supported:

Behavior aggregate (BA) or code point traffic classifiers—Code points determine each packet's forwarding class and loss priority. BA classifiers allow you to set the forwarding class and loss priority of a packet based on DiffServ code point (DSCP) bits, IP precedence bits, MPLS EXP bits, and IEEE 802.1p bits. The default classifier is based on IP precedence bits.

Multifield (MF) traffic classifiers—Allow you to set the forwarding class and loss priority of a packet based on firewall filter rules. For more information about configuring MF classifiers, see the *JUNOS Internet Software Configuration Guide: Policy Framework*.

Forwarding classes—Also known as ordered aggregates in the IETF's DiffServ architecture. Affect the forwarding, scheduling, and marking policies applied to packets as they transit a router. Four forwarding classes are supported: best effort, assured forwarding, expedited forwarding, and network control. The forwarding class plus the loss priority define the per-hop behavior.

Loss priorities—Allow you to set the priority of dropping a packet. Typically you mark packets exceeding some service level with a high loss priority. Loss priority affects the scheduling of a packet without affecting the packet's relative ordering. You set loss priority by configuring a classifier or a policer.

Forwarding policy options—Allow you to associate forwarding classes with next hops. Forwarding policy also allows you to create classification overrides, which assign forwarding classes to sets of prefixes.

Transmission scheduling and rate control—Provide you with a variety of tools to manage traffic flows:

Schedulers—Allow you to define the priority, bandwidth, delay buffer size, rate control status, and RED drop profiles to be applied to a particular forwarding class for packet transmission.

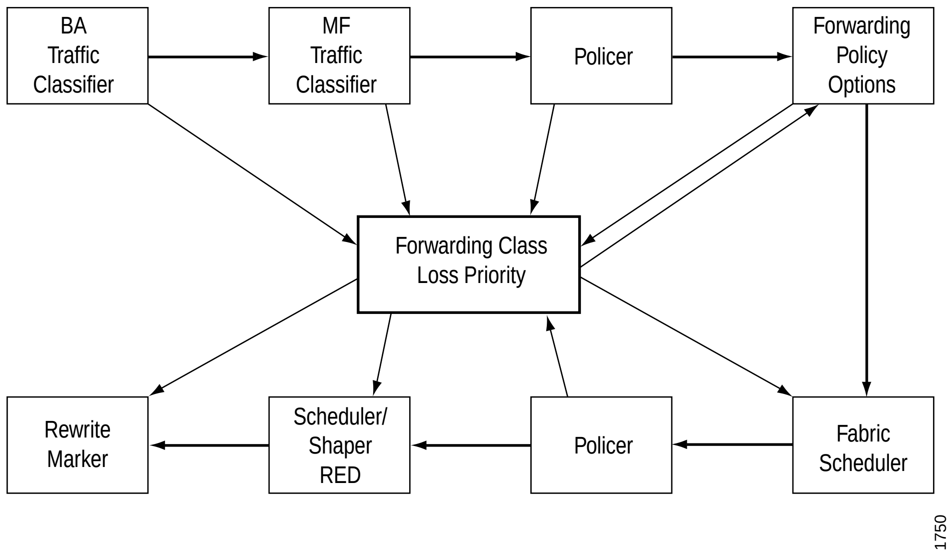
Fabric schedulers—For T-series platforms only, fabric schedulers allow you to identify a packet as high or low priority based on its forwarding class.

Policers—Allow you to limit traffic of a certain class to a specified bandwidth and burst size. Packets exceeding the policer limits can be discarded, or can be assigned to a different forwarding class or to a different loss priority, or to both. You define policers with filters that can be associated with either input or output interfaces. For information about configuring policers, see the *JUNOS Internet Software Configuration Guide: Policy Framework*.

Rewrite markers—Allow you to redefine the code-point value of outgoing packets. Rewriting or *marking* outbound packets is useful when the router is at the border of a network and must alter the code points to meet the policies of the targeted peer.

Figure 22 shows the components of the JUNOS CoS features, illustrating the sequence in which they interact.

Figure 22: Packet Flow Through CoS Configurable Components



The components are discussed in the following sections:

Traffic Classifiers on page 412

Forwarding Classes on page 414

Transmission Scheduling and Rate Control on page 415

Rewrite Markers on page 417

Traffic Classifiers

By default, all logical interfaces are assigned an IP precedence *classifier* for incoming IP packets.

At the core router, the software matches the classifier to a *code point* to determine each packet's forwarding class and loss priority. This classifier is called the behavior aggregate (BA) classifier. Supported code points include the DiffServ Code Point (DSCP) for IP DiffServ, IP precedence bits, MPLS EXP bits, and IEEE 802.1p CoS bits.

In an edge router, a multifield (MF) classifier provides the filtering functionality that scans through a variety of packet fields to determine the forwarding class for a packet. Typically, a classifier performs matching operations on the selected fields against a configured value.

There are only four separate classes that can forward traffic independently. Therefore, you must configure additional classes to be aggregated into one of these four classes. You configure class aggregation using the BA classifier. For more information, see "Forwarding Classes" on page 414.

The following sections discuss classifiers in more detail:

Default Classifier

When you install a classifier, it becomes effective on any interface for which you configure it.

By default, all logical interfaces are assigned an IP precedence classifier. The default IP precedence classifier maps IP precedence bits to forwarding classes and loss priorities as shown in Table 22.

Table 22: Default IP Precedence Classifier

IP Precedence Code Point	Forwarding Class	Loss Priority
000	best-effort	low
001	best-effort	high
010	best-effort	low
011	best-effort	high
100	best-effort	low
101	best-effort	high
110	network-control	low
111	network-control	high

Behavior Aggregate Classifier

A behavior aggregate classifier uses IP DSCPs, IP precedence bits, the MPLS EXP field, or Layer 2 CoS indication (IEEE 802.1p) to determine the forwarding treatment for each packet, called a per-hop behavior (PHB). A PHB defines how a particular router in a DiffServ domain treats a packet. A BA classifier can aggregate multiple DiffServ PHBs into a single one if the router cannot support multiple simultaneous PHBs.

The BA classifier maps a code point to a loss priority. The loss priority is used later in the work flow to select one of the two drop profiles employed by RED.

Decoding the EXP header field can also determine the packet loss priority (PLP) status.

> **Note**
> For a specified interface, you can configure both an MF classifier and a BA classifier without conflicts. Because the classifiers are always applied in sequential order, the BA classifier followed by the MF classifier, any BA classification result is overridden by an MF classifier if they conflict.

For information about configuring BA classifiers, see “Classify Packets by Behavior Aggregate” on page 427 and “Example: Configure Class of Service” on page 436.

Multifield Classifier

A multifield classifier examines one or more packet fields to determine the forwarding treatment that a packet receives. An MF classifier typically matches one or more of the six packet header fields: destination address, source address, IP protocol, source port, destination port, and DSCP. MF classifiers are used when a simple BA classifier is insufficient to classify a packet.

From a CoS perspective, MF classifiers (or firewall filter rules) provide the following services:

- Classify packets to a forwarding class and loss priority.

- Police traffic to a specific bandwidth and burst size. Packets exceeding the policer limits can be discarded, or can be assigned to a different forwarding class or to a different loss priority, or to both.

To activate an MF classifier, you must configure it on a logical interface. There is no restriction on the number of MF classifiers you can configure.

For information about configuring MF classifiers, see the *JUNOS Internet Software Configuration Guide: Policy Framework*.

Forwarding Classes

For a classifier to assign an output queue to each packet, it must associate the packet with one of the following forwarding classes:

- Expedited Forwarding (EF)—Provides a low loss, low latency, low jitter, assured bandwidth, end-to-end service.

- Assured Forwarding (AF)—Provides a group of values you can define and includes four subclasses, AF1, AF2, AF3, and AF4, each with three drop probabilities, low, medium, and high.

- Best Effort (BE)—Provides no service profile. For the BE forwarding class, loss priority is typically not carried in a code point and RED drop profiles are more aggressive.

- Network Control (NC)—The NC forwarding class is typically low priority because it is adaptive.

For each forwarding class, you can configure high or low loss priority. By default, the loss priority is low. For information about configuring forwarding classes, see "Configure Forwarding Classes" on page 425 and "Example: Configure Class of Service" on page 436.

Transmission Scheduling and Rate Control

You configure transmission scheduling and rate control parameters using *scheduler s*. Schedulers define the priority, bandwidth, delay buffer size, rate control status, and RED drop profiles to be applied to a particular class of traffic.

You associate the schedulers with forwarding classes by means of *scheduler maps* . You can then associate each scheduler map with an interface, thereby configuring the hardware queues, packet schedulers, and RED processes that operate according to this mapping.

The following sections describe these processes in more detail:

Scheduling Priority

The scheduling priority determines the order of transmission from the forwarding classes associated with an output interface. Three levels of transmission priority are currently supported: low, high, and strictly high.

High-priority forwarding classes transmit packets ahead of low-priority forwarding classes as long as the forwarding class retains enough bandwidth credit. When you configure a high-priority forwarding class with a significant fraction of the transmission bandwidth, the forwarding class might lock out low-priority traffic.

Strictly high-priority forwarding classes receive precedence over low-priority forwarding classes as long as the forwarding class has traffic waiting to be sent, irrespective of bandwidth credit. We recommend that you do not configure strictly high and high transmission priorities on a single interface, unless the interface sends network-control traffic with a need for 5 percent of the transmission bandwidth.

Strictly high-priority forwarding classes supersede bandwidth guarantees for low-priority forwarding classes; therefore, we recommend that you use this feature to ensure proper ordering of special traffic, such as voice traffic. You can preserve bandwidth guarantees for low-priority forwarding classes by allocating to the strictly high-priority forwarding class only the amount of bandwidth that you generally require for that forwarding class. For example, consider the following allocation of transmission bandwidth:

Q0 BE—20 percent, low priority

Q1 EF—30 percent, strictly high priority

Q2 AF—40 percent, low priority

Q3 NC—10 percent, low priority

This allocation of bandwidth assumes that, in general, the EF forwarding class requires only 30 percent of an interface's transmission bandwidth. However, if short bursts of traffic are received on the EF forwarding class, 100 percent of the bandwidth is given to the EF forwarding class by way of the strictly high-priority setting.

For information about configuring scheduling priority, see "Configure Scheduling Policy Maps" on page 429 and "Example: Configure Class of Service" on page 436.

Fabric Priority Queuing

On T-series platforms, the default behavior is for fabric priority queuing on egress interfaces to match the scheduling priority you assign. High-priority egress traffic is automatically assigned to high-priority fabric queues. Likewise, low-priority egress traffic is automatically assigned to low-priority fabric queues.

For information about overriding automatic fabric priority queuing, see "Configure Forwarding Classes" on page 425.

Transmission Rate Control

The transmission rate control determines the actual traffic bandwidth from each of the forwarding classes you configure. The rate is specified in bits per second. You can limit the transmission bandwidth to the exact value you configure, or allow it to exceed the configured rate if additional bandwidth is available from other queues.

For information about configuring transmission rate control, see "Configure Scheduling Policy Maps" on page 429.

Allocation of Leftover Bandwidth

When a forwarding class fails to fully use the allocated transmission bandwidth, the remaining bandwidth can be taken by other forwarding classes if they receive a larger amount of offered load than the bandwidth allocated. This use of leftover bandwidth is the default behavior. If you want a forwarding class to not take any extra bandwidth, you must configure it for strict allocation. With rate control in place, the specified bandwidth is strictly observed.

When you configure more than one forwarding class to use leftover bandwidth, the high-priority forwarding class takes the bandwidth first. Forwarding classes with equal priority share the bandwidth through round robin.

For information about configuring leftover bandwidth allocation, see "Configure Scheduling Policy Maps" on page 429.

Default Congestion and Transmission Control

A default congestion and transmission control mechanism is needed when an output interface is not configured for a certain forwarding class, but receives packets destined for that unconfigured forwarding class. This default mechanism uses the delay buffer and WRR credit allocated to the designated forwarding class, with a default drop profile. Because the buffer and WRR credit allocation is minimal, packets might be lost if a larger number of packets are forwarded without configuring the forwarding class for the interface.

RED Congestion Control

You can configure two parameters to control congestion at the output stage. The first parameter defines the delay-buffer bandwidth, which provides packet buffer space to absorb burst traffic up to the specified duration of delay. Once the specified delay buffer becomes full, packets with 100% drop probability are dropped from the head of the buffer.

The second parameter defines the drop probabilities across the range of delay-buffer occupancy, supporting the RED process. Depending on the drop probabilities, RED might drop packets aggressively long before the buffer becomes full, or it might drop only a few packets even if the buffer is almost full.

You specify the delay-buffer size for each scheduler associated with an output interface configuration in units of milliseconds, or as a percentage of the entire interface buffer space. You specify drop probabilities in the drop profile section of the CoS configuration hierarchy and reference them in each scheduler configuration. For each scheduler, you can configure four separate drop profiles, one for each combination of loss priority (low or high) and IP transport protocol (TCP or non-TCP).

You can configure a maximum of 32 different drop profiles.

For information about configuring delay buffers and drop profiles, see “Configure Scheduling Policy Maps” on page 429 and “Configure RED Drop Profiles” on page 430.

Rewrite Markers

A marker reads the current forwarding class and loss priority information associated with a packet and finds the chosen code point from a table. It then writes the code point information into the packet header. Entries in a marker configuration represent the mapping of the current forwarding class into a new forwarding class, to be written into the header.

You define markers in the rewrite rules section of the CoS configuration hierarchy and reference them in the logical interface configuration. When an interface is not associated with any marker, the code point is not rewritten; instead, the old marking information is preserved.

This model supports marking on the IP ToS byte, the MPLS EXP bits, and Layer 2 IEEE 802.1p CoS indications.

For information about configuring rewrite markers, see “Rewrite Packet Header Information” on page 431 and “Example: Configure Class of Service” on page 436.

Hardware Capabilities and Limitations

Juniper Networks T-series platforms and M-series platforms with enhanced FPCs can use an expanded range of CoS capabilities as compared to M-series platforms that employ the earlier FPC model. Table 23 on page 421 lists these differences between the original FPCs and the Enhanced FPCs.